INTISARI

Pengembangan *Word Embedding* untuk Domain Spesifik Ulasan Hotel Berbahasa Indonesia

Oleh

NADIA ELAESIANA PUTRI
16/398523/PA/17484

Word embedding merupakan salah satu metode representasi kata. *Word embedding* mempunyai kelebihan dibanding representasi kata lain, yaitu dapat menangkap hubungan semantik pada kata. *Pre-trained word embedding* merupakan *word embedding* yang dilatih pada korpus yang besar dan domain yang umum. *Pre-trained word embedding* menghasilkan performa yang terbatas dalam menyelesaikan masalah NLP pada domain spesifik karena tidak menangkap informasi semantik pada domain spesifik.

Pada penelitian ini akan dikembangkan domain spesifik *word embedding* untuk ulasan hotel berbahasa Indonesia. Domain spesifik *word embedding* dievaluasi secara intrinsik dan ekstrinsik. Evaluasi intrinsik dilakukan dengan melihat kemampuan model domain spesifik *word embedding* dalam menangkap sinonim dan kata turunan. Evaluasi ekstrinsik yang dilakukan adalah analisis sentimen. Evaluasi juga dilakukan pada *pre-trained word embedding* dan kombinasi *pre-trained* dan domain spesifik *word embedding*. Model *word embedding* yang dibuat pada penelitian ini adalah Word2Vec dan FastText.

Pada evaluasi intrinsik, domain spesifik *word embedding* memperoleh akurasi lebih rendah dibanding *pre-trained word embedding*, yaitu akurasi domain spesifik *word embedding* untuk Word2Vec dan FastText sebesar 0.56 dan akurasi *pre-trained word embedding* sebesar 0.65 untuk Word2Vec dan 0.62 untuk FastText. Pada evaluasi ekstrinsik, domain spesifik *word embedding* memiliki performa paling tinggi, yaitu memperoleh akurasi 0.76, *recall* 0.76, dan *F1-Score* 0.79. Sedangkan *pre-trained word embedding* memiliki performa lebih rendah dibanding domain spesifik *word embedding*, terutama model *pre-trained word embedding* FastText yang memiliki performa paling rendah diantara semua model, yaitu dengan akurasi 0.71, *recall* 0.71, dan *F1-Score* 0.75.

Kata-kata kunci : *word embedding*, Word2Vec, FastText, analisis sentimen.

ABSTRACT

Development of Domain Specific Word Embedding for Indonesian Hotel Reviews

By

NADIA ELAESIANA PUTRI
16/398523/PA/17484

Word embedding is one of the word representation methods. One of the advantages of word embedding compared to other word representation methods is that it can capture semantic relation between the words. Pre-trained word embedding is word embedding trained on large-scale generic corpora. The performance of pre-trained word embedding for solving NLP tasks in domain specific is limited, since pre-trained word embedding do not capture domain specific semantics/knowledge.

In this study, domain specific word embedding will be developed with Indonesian hotel reviews as the domain. Intrinsic and extrinsic evaluation will be conducted to evaluate domain specific word embedding. Intrinsic evaluation will be done by measuring the ability of domain specific word embedding in capturing synonyms and alternative forms. Extrinsic evaluation will be done with sentiment analysis. Evaluations are also carried out on pre-trained word embedding and combination of pre-trained and domain-specific word embedding. Word embedding models that will be built in this study are Word2Vec and FastText.

In intrinsic evaluation, domain specific word embedding obtained lower accuracy than pre-trained word embedding, with accuracy of 0.56 for Word2Vec and FastText domain specific word embedding, and accuracy of 0.65 for Word2Vec pre-trained word embedding and 0.62 for FastText pre-trained word embedding. In extrinsic evaluation, domain specific word embedding has the highest performance, with an accuracy of 0.76, *recall* 0.76, and *F1-Score* 0.79. Whereas pre-trained word embedding has lower performance than domain specific word embedding, especially FastText pre-trained word embedding which has the lowest performance among all models, with an accuracy of 0.71, *recall* 0.71, and *F1-Score* 0.75.

Keywords : word embedding, Word2Vec, FastText, sentiment analysis.